Variance Reduction for Error Estimation When Classifying Colon Polyps from CT Colonography

James D. Malley*[a], Anna K. Jerebko[b],
Meghan T. Miller[b], Ronald M. Summers[b]

[a] Center for Information Technology, NIH, Bethesda MD 20892
[b] Clinical Center, Department of Diagnostic Radiology,
NIH, Bethesda MD 20892

ABSTRACT

For cancer polyp detection based on CT colonography we investigate the sample variance of two methods for estimating the sensitivity and specificity. The goal is the reduction of sample variance for both error estimates, as a first step towards comparison with other detection schemes. Our detection scheme is based on a committee of support vector machines. The two estimates of sensitivity and specificity studied here are a smoothed bootstrap (the 632+ estimator), and ten-fold cross-validation. It is shown that the 632+ estimator generally has lower sample variance than the usual cross-validation estimator. When the number of nonpolyps in the training set is relatively small we obtain approximately 80% sensitivity and 50% specificity (for either method). On the other hand, when the number of nonpolyps in the training set is relatively large, estimated sensitivity (for either method) drops considerably. Finally, we consider the intertwined roles of relative sample sizes (polyp/nonpolyp), misclassification costs, and bias-variance reduction.

Keywords: virtual colonoscopy, classification, error estimation, support vector machines, bias-variance reduction

1. INTRODUCTION

In any classification problem it is important to obtain error estimates for the classifier that have both low bias and low variance, in addition to constructing a classifier that has good sensitivity and specificity. A standard method for obtaining these error estimates with generally good statistical properties is cross-validation, typically with k (= number of folds) ten or higher. In this study we look at an alternative to error estimation scheme that appears to have lower sample variance than cross-validation, namely the smoothed bootstrap 632+ method introduced by Efron & Tibshirani, 1997. We do this in the context of a decision engine based on a committee of support vector machines applied to the problem of detecting colon cancer polyps. The full feature set is derived from a 3D reconstruction derived from a CT colonography. Subsets of features are identified using a genetic algorithm (GA), and a simple majority vote is made across the committee of SVMs each trained using a different feature subset. The true colon polyps are those detected by a complete colonoscopy, our gold standard. The nonpolyps are those declared

* jmalley@helix.nih.gov phone 301 496 9934; fax 301 402 2867

Medical Imaging 2003: Physiology and Function: Methods, Systems, and Applications, Anne V. Clough, Amir A. Amini, Editors, Proceedings of SPIE Vol. 5031 (2003) © 2003 SPIE · 1605-7422/03/$15.00

to be (possible) polyps after a series of thresholds and simple filters is applied to the CT reconstructions.

We begin with a description of the dataset and the features used for detection, discuss the structure of the support vector machines (SVMs) used for classification in our committee approach, briefly describe the smoothed-bootstrap 632+ estimator (Efron & Tibshirani, 1997), and provide results of numerical experiments showing the variability of 632+ and 10xCV (for both sensitivity and specificity) as the number of nonpolyps in the training dataset is adjusted.

Finally, we briefly discuss approximating the optimal value of the relative number of nonpolyps to true polyps in the training set needed to generate high sensitivity rates that also have low sample variance. We observe that adjusting the number of nonpolyps in the training set is partly equivalent to adjusting the misclassification costs in training the decision engine. We point out that an unintended consequence of adjusting the number of nonpolyps used in training (to increase sensitivity and optimally reduced sample variance) is that such adjustment implies misclassification weights that may not correspond to those understood or desired by the user of the decision engine.

2. MATERIALS

The data set contained 80 studies consisting of supine and prone screening of 40 average risk patients half of whom had at least one 1 cm or larger polyp and the remaining 20 had no polyps. The 20 patients with polyps typically had polyps smaller than 1 cm also. CT scans were done on G.E.Lightspeed scanners. Scanning parameters were 120kVp, 50mAs (mean), field of view to fit (38-46 cm), 5 mm collimation, HQ mode, and 3 mm reconstruction interval (2mm overlap). CT images were processed to three-dimensional surface renderings of the colon by our research software. All 40 patients underwent complete colonoscopy examination. Among the 20 cases there were 65 polyps, of which 25 were large polyps (1 cm or larger), and of these only 18 are identified by the radiologist.

The feature set used is derived from software that first segments the colon using a region growing algorithm. Regions of interest along the colon wall are identified. A total of 80 quantitative features are defined for each polyp candidate, but we have seen that not all of these features are eventually useful. Useful features include: maximum average polyp neck geometric curvature, wall thickness, polyp volume, and average volumetric Gaussian curvature. Further details can be found in Summers et al. (2002), and Miller et al. (2003).

3. METHODS

Our basic classification scheme is a *support vector machine* (SVM) and our proposed alternative to the 10xCV method for error estimation is the 632+ estimator (Efron & Tibshirani, 1997). We first outline the structure of a SVM, then discuss the 632+ scheme, outline our feature selection approach, and conclude with a discussion of the error estimates, misclassification costs and sample sizes.

3.1 Support Vector Machines (SVMs) An essential reference is Hastie et al. (2001); additional material can be found in Cristianini & Shawe-Taylor (2000) and Schölkopf et al. (eds.) (1999). Consider first constructing an SVM based on the original data, which

consists of N pairs $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$, for p-dimensional features (predictors) x_i and outcomes (polyp, nonpolyp) $y_i = +1, \text{or } -1$.

Define a linear decision boundary (hyperplane) by

$$\{x : f(x) = x^T\beta + \beta_0 = 0\},$$

where β is a unit vector $\|\beta\| = 1$. We define a classification rule based on *f(x)* by

$$G(x) = \text{sign}[f(x)]$$

such that for the test pair (x,y), the observation y is declared a polyp if $G(x) > 0$, and a non-polyp if $G(x) \leq 0$. We observe that *f(x)* is the signed distance from the data point x to the hyperplane defined by $f(x) = 0$. We define the *margin* C for the SVM to be such that $2C = 2/\|\beta\|$.

Optimal estimation of $\{\beta, \beta_0\}$ is quadratic with linear inequality constraints, and thus is a convex optimization problem. When the two classes $\{y = +1\}, \{y = -1\}$ can be separated by a hyperplane in the feature space, then the optimization can be re-expressed as:

$$\min\|\beta\| \quad \text{for all } \beta, \beta_0$$

subject to

$$y_i(x_i^T\beta + \beta_0) \geq 1, \quad i = 1, ..., N.$$

More generally, when the two classes cannot be separated (the usual case), then define so-called *slack variables* $\xi = (\xi_1, \xi_2, ..., \xi_N)$, and modify the constraints to

$$y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \quad \text{for all } i,$$

where

$$\xi_i \geq 0 \quad \text{and} \quad \Sigma\xi_i \leq \text{ constant}.$$

The general solution for β has the form

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i,$$

for coefficients $\hat{\alpha}_i$ subject to the constraints

$$0 \leq \alpha_i \leq \gamma, \qquad \sum_{i=1}^{N} \alpha_i y_i = 0,$$

and

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] = 0,$$

$$\mu_i\xi_i = 0,$$

$$y_i(x_i^T\beta + \beta_0) - (1 - \xi_i) \geq 0,$$

for $i = 1, 2, ..., N$.

In the above we have defined

$$\alpha_i = \gamma - \mu_i, \quad \text{for all } i$$

where γ is a tuning parameter, set by the user.

It can be seen that the solution $\hat{\beta}$ has nonzero coefficients $\hat{\alpha}_i$ only for those observations i for which the constraints

$$y_i(x_i^T\beta + \beta_0) - (1 - \xi_i) \geq 0$$

are exactly met (equality obtains). Such observations are called *support vectors*. These data points are of two types: those that lie on the wrong side of the decision boundary,

and those that lie on the correct side of the boundary but that are close to it, that is, they reside within the margin.

To place the SVM scheme in a larger context, we note that the optimization problem can be re-stated as a *penalized likelihood* problem, such that $f(x) = x^T\beta + \beta_0$ solves the problem

$$\min_{\beta,\beta_0} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \lambda \|\beta\|^2,$$

and where $\lambda = 1/2\gamma$, and the subscript "+" denotes the positive part of the function. This has the general form of *loss function + penalty function.*

The optimization problem also an elegant statement in terms *reproducing kernel Hilbert spaces* (see Hastie et al., 2001; pp. 377-384), and the mathematics of the subject is rich and well studied. In this context the penalty function above can be generalized, and many choices are then available for the kernel K, among which are polynomials of user-specified degree d, radial basis functions, or weighted hyperbolic tangents (so-called neural network kernels). Using alternative loss functions leads to different classification schemes: the binomial log-likelihood generates the logistic regression scheme, and squared-error loss leads to a penalized linear discriminant decision rule; see Hastie et al., (1997), p. 381.

An important further extension of the SVM architecture described above is the use of functions, $h_j(x)$, $j = 1, 2, \ldots, M$, of the original data vector x. It is possible that such functions transform the problem into a nearly linear one in a sufficiently high dimensional space, and thus that the decision boundary can be easily found, and possibly such that the data classes are fully separated. The search for such superior separating functions may be unrewarding however, and we expect instead to see that a large family of transformations produces closely similar results.

We now consider the error estimate problem and the possible reduction in sample variance of the estimates.

3.2 The 632+ error estimator It is well known that for any classification problem the observed (or, apparent) error rate, derived from testing a decision engine on the same data as it was trained on, routinely gives estimates that are much too optimistic; see for example Efron (1986).

Ideally we want error estimates (of sensitivity and specificity) to satisfy three criteria at once: high mean values, low bias, and low variance. Some of the alternatives to the apparent error estimate have low bias or low variance, but often not both at once. The first requirement of high mean values is, in our perspective, a function of finding a good decision engine and is properly a model selection problem. Here we work with decision engines that are ensemble versions of SVMs, where it has been observed that ensemble or committee versions of any reasonable decision engine tend to improve error rates over any single engine in the same class; see Dietterich (1999).

As alternatives to apparent error rate, procedures based on cross-validation and the bootstrap have been proposed, and extensively studied; see Efron & Tibshirani (1997). Constructing the algorithm begins as follows:

Let N be the size of the training dataset X

1) Set aside a single case in the training data, say (x_1, y_1)
2) Of the remaining *N-1* cases, draw a bootstrap sample X^* of size N

3) Train the classifier on X^* and make a prediction for the case (x_1, y_1)
4) Repeat Steps (2) and (3) K times, each time making a prediction for (x_1, y_1)
5) Replace (x_1, y_1) in the dataset and consider the next case (x_2, y_2)
6) Repeat Steps (2) and (3) and make a prediction for the case (x_2, y_2)
7) Repeat Step (6) for all cases in the dataset;
8) The average error rate over all predictions is called $Err^{(1)}$

Efron & Tibshirani refer to $Err^{(1)}$ as the (simple) bootstrap smoothed leave-one-out estimator, and observe that while it has desirably lower variance than the usual leave-one-out estimator (involving no resampling and thus no smoothing), it has a noticeable bias upward: it yields a pessimistic outcome, as opposed to the apparent error rate, $\overline{err}$ that, as noted, is uniformly too optimistic. To adjust for this bias, they introduce two further innovations that attempt to reduce the bias and still maintain low variance. Thus, let

$$Err^{(632)} = .368(\overline{err}) + .632(Err^{(1)}).$$

The weights are derived from the fact that a bootstrap sample in general captures only the fraction .632 of any set of cases, and so the bootstrapped estimator $Err^{(1)}$ needs to be weighted relative to $\overline{err}$ Since the estimator $\overline{err}$ is biased downward, it is presumed that $Err^{(632)}$ will be less biased upward.

However, with the bias adjustment above comes the likelihood that the low variance property of $Err^{(1)}$ will have been disturbed, and the simulations in Efron & Tibshirani confirm this. Hence they introduce a second adjustment, aimed at reducing the overfitting inherent in $\overline{err}$ The new estimator, called $Err^{(.632+)}$, appears from their simulations to have reduced sample variance: see Efron & Tibshirani (1997) for full details and motivation for the adjustment. Briefly, for the decision rule $r_X(y)$ define the loss function

$$Q[y, r_X(y)] = +1 \text{ if } r_X(y) = y, \quad = 0 \text{ otherwise.}$$

Then let

$$\hat{\gamma} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} Q[y_i, r_X(y_j)].$$

The estimator $\hat{\gamma}$ is intended to capture the no information error that would result if the data X and the class variable y were statistically independent: the predictor X has no information concerning the class variable y. In such a case $\overline{err}$ would be .50, for any classification rule.

Next, define

$$Err^{(1)\prime} = \min(Err^{(1)}, \hat{\gamma})$$

and

$$\hat{R}' = (Err^{(1)\prime} - \overline{err})/(\hat{\gamma} - \overline{err}) \text{ if } Err^{(1)} \text{ and } \hat{\gamma} > \overline{err}$$
$$= 0 \text{ otherwise.}$$

Finally, we arrive at

$$Err^{(.632+)} = Err^{(.632)} + (Err^{(1)\prime} - \overline{err}) \cdot \frac{.368 \cdot .632 \cdot \hat{R}'}{1 - .368\hat{R}'}.$$

The series of adjustment made above, though largely heuristic, were reasonably successful in the numerous simulations discussed in Efron & Tibshirani (1997).

3.3 Feature Selection using a Genetic Algorithm (GA) In earlier work it has been found that using smaller sets of features for training classifiers for colon polyps detection

leads to better detection results (lower bias) and reduced variance; see Jerebko et al. (2003). Thus a neural net classifier having many input features tends to do less well than one with fewer features, when applied to test data. This is a generally observed result, and relates to the trade-off between model complexity and generalizability (accuracy on test data); see, for example Hastie et al. (2001) Chapter 7.

Thus, we sought to find small sets of features (here, four) as input to a committee classifier, and to obtain good sets we implemented a *genetic algorithm*. This returned sets of features that, separately, were good predictors when used as input to an SVM; see Miller, et al (2003) for details. The fitness function used therein was the 632+ estimator studied here. In this study of feature selection it was found that the GA approach uncovered sets of features that had approximately the same sensitivity, but markedly improved specificity, when compared to a simple forward stepwise selection approach.

3.4 The Committee Approach to CAD for Colon Polyps Use of a committee (or, ensemble) approach to classification has been the object of much recent research in the machine learning literature. The committee used here requires construction of a number of base classifiers (SVMs), followed by a simple majority vote across the component SVMs. Other ensemble methods are being studied construct the base classifiers in different ways or use iterative weighting across the chosen classifiers, and such committees include methods such as *bagging*, *boosting*, and the *general additive model*; see for example Dietterich (2002). Of particular interest is the method of *logistic boosting*, which returns not a simple class assignment for each case, but a probability class estimate for the case. These more involved methods are currently being evaluated for our colon polyp detection data.

4. RESULTS

We being by forming a committee (using majority voting) of six SVMs, each based on features selected by the GA algorithm; see Table 1 for the feature list. The committee classifier was then applied to the colonography dataset described above, and both 10xCV and 632+ were used in estimating sensitivity and specificity. This process was repeated 100 times. Figures 1 – 8 display the results of our numerical experiments. ROC curves are displayed wherein the number of nonpolyps used in the training set was varied. We compared the usual 10xCV method for error estimation with the 632+ estimator discussed above; standard deviations (std) across the 100 runs are given (standard errors = std/10). We made comparisons for both methods when the cost of misclassification was varied, and also display results for varying the nonpolyp sample size used in training the committee of SVMs.

Figs. 1 and 2 show the outcomes for 10xCV, with equal cost for misclassification. We find that a reasonable choice for the nonpolyp sample size is approximately 20, at which point the sensitivity is 80% with a specificity of 60%. However, the std of the sensitivity estimate at this point is rather high (0.2). Figs. 3 and 4 show the results of using a weight for misclassification (true polyp/nonpolyp = 2.5). Here a reasonable choice of nonpolyp sample size is approximately 40, at which the sensitivity is again 80% with a specificity of 60%. However, we see that the variability of the sensitivity estimate has dropped considerably (now std = 0.05) while the std for specificity has increased slightly.

Review of the corresponding plots for the 632+ estimator (Figs. 5, 6, 7, and 8) reveals a similar story regarding the optimal nonpolyp sample size in the training set. However, they also reveal a difference between 10xCV and 632+ as error estimators: the 632+ estimator has reduced variance relative to 10xCV when the number of nonpolyps is above 50, for both the equal and weighted costs. On the other hand, the 632+ estimator has greatly reduced, and unacceptable, sensitivity for these higher nonpolyp sample sizes: below 40% sensitivity for both equal and weighted cost.

CONCLUSIONS

Our proposed estimator, 632+, can indeed reduce variance in the estimation of sensitivity and specificity (compared to 10xCV), but does so only when the sample size of the nonpolyps in the training set is sufficiently large (above 20 with our dataset). However, in this range the observed average sensitivity (using either 632+ or 10xCV) is unacceptably low: drops below 80% with equal costs. With unequal costs (true/nonpolyp = 2.5) sensitivity remains above 80% with 40 nonpolyps in the training set, and the 632+ variance remains consistently below that of 10xCV.

We conclude that optimal sensitivity and specificity (above 80%, and 50% respectively) was found with approximately 20 nonpolyps in the training set, and that for such data the 632+ estimator had reduced variability relative to10xCV when using equal costs. When unequal costs are imposed, optimal sensitivity and specificity was found with approximately 40 nonpolyps in the training set, and here again the 632+ estimator had reduced variability relative to 10xCV. A more complete study of cost functions and sample sizes is required to draw conclusions regarding optimal classifying engines and datasets, but in all cases considered here 632+ generally improves on 10xCV for obtaining error estimates with reduced variance.

REFERENCES

N Cristianini, J Shawe-Taylor, *Support Vector Machines, and other Kernel-Based Learning Methods*. Cambridge University Press; 2000.

TG Dietterich, "Ensemble Learning." In *The Handbook of Brain Theory and Neural Networks, Second edition,* (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002.

TG Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization." *Machine Learning*, 1-22; 1999

B Efron, "How biased is the apparent error rate of a prediction rule?" *Journal of the American Statistical Association*, **81(394)**, 461-470. 1986.

B Efron, R Tibshirani, "Improvements on cross-validation: the .632+ bootstrap method." *Journal of the American Statistical Association*, **92(438)**, 548-560. 1997.

T Hastie, R Tibshirani, J Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York; Springer; 2001.

AK Jerebko, JD Malley, M Franaszek, RM Summers, "Multi-network classification scheme for detection of colonic polyps in CT colonography data sets." In press: *Academic Radiology*; 2003.

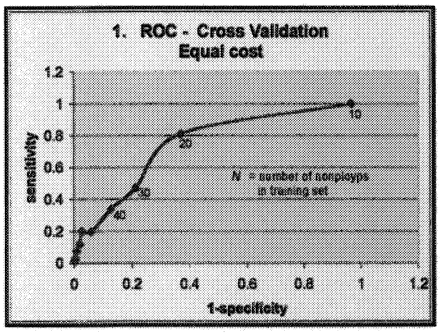
MT Miller, AK Jerebko, JD Malley, RM Summers, "Optimization of the support vector classifier using genetic algorithms for variable selection." *Proceedings SPIE Medical Imaging Conference*; 2003.

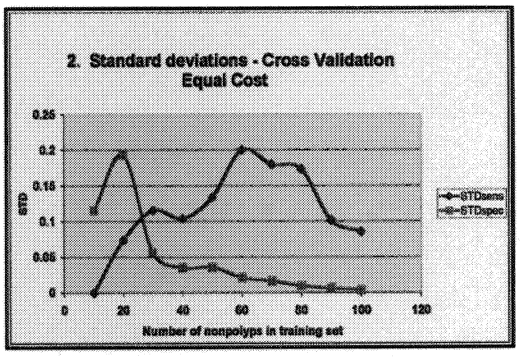
B Schölkopf, C Burges, AJ Smola (eds), *Advances in Kernel Methods; Support Vector Learning*. MIT Press; 1999.

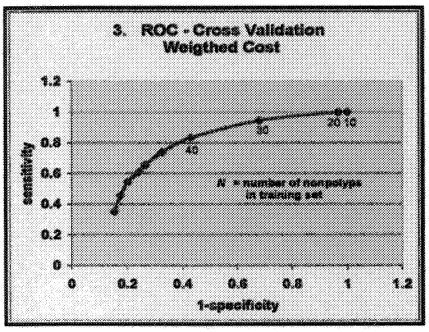
RM Summers, AK Jerebko, M Franaszek, JD Malley, CD Johnson, "Colonic polyps: complementary role of computer-aided detection in CT colonography." *Radiology*, **225**, 391-399.

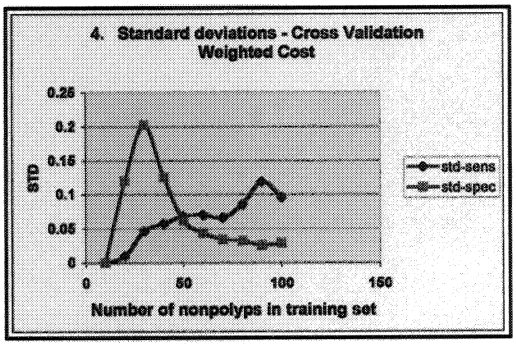
R Tibshirani, "A comparison of some error estimates for neural network models." Technical Report, Stanford University Department of Statistics; 1995.

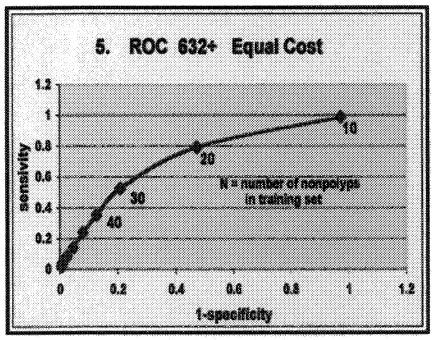
G Valentini, TG Dietterich, "Bias-Variance analysis and ensembles of SVM." In J. Kittler and F. Roli (Ed.) *Third International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*. New York: Springer Verlag; 2002.

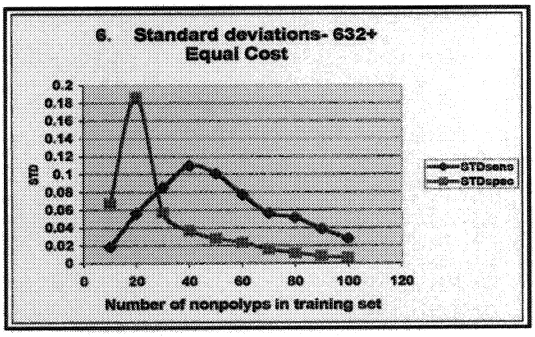

1. ROC - Cross Validation
Equal cost
sensitivity
1-specificity
N = number of nonpolyps in training set
10
20
30
40

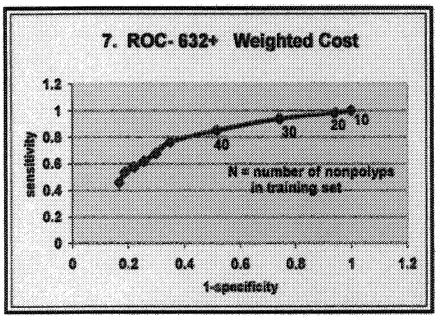

2. Standard deviations - Cross Validation
Equal Cost
STD
Number of nonpolyps in training set
STDsens
STDspec

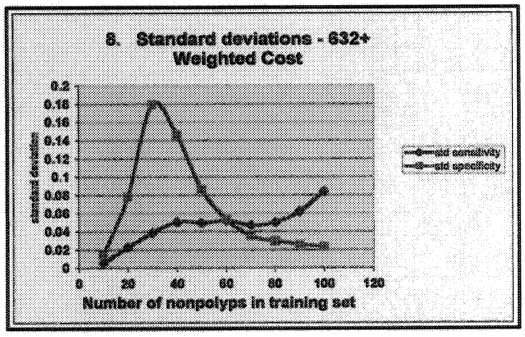

3. ROC - Cross Validation
Weigthed Cost
sensitivity
1-specificity
N = number of nonpolyps in training set
10
20
30
40


4. Standard deviations - Cross Validation
Weighted Cost
STD
Number of nonpolyps in training set
std-sens
std-spec


5. ROC 632+ Equal Cost
sensitivity
1-specificity
N = number of nonpolyps in training set
10
20
30
40


6. Standard deviations- 632+
Equal Cost
STD
Number of nonpolyps in training set
STDsens
STDspec


7. ROC- 632+ Weighted Cost
sensitivity
1-specificity
N = number of nonpolyps in training set
10
20
30
40


8. Standard deviations - 632+
Weighted Cost
standard deviation
Number of nonpolyps in training set
std sensitivity
std specificity